Report for assignment 3 of COMP90049

Anonymous

1 Introduction

This report examines and analyses several feature engineering methods, models, and model ensemble methods on count dataset. In the feature engineering process, averaging the term frequency of tweets from the same users and selecting features with xgboost and chi2 are proved to be helpful to improve the performance. Besides, the feature selecting results suggests that the region-related lexical variation are the most effective features in distinguishing user locations. Regarding the model selection, the xgboost and bagged xgboost have the highest accuracy and f1 score among all the tested models. Another finding is the uneven classes may cause serious biases for most models, which result to one region's absent from the prediction.

2 Review

This study is mainly carried on a dataset including users, their region and the count of pre-selected words in their tweets. Apart from that, the text tweet dataset including users, regions and their raw tweets are also used in this report to train the bert model. Those data are generated from the Twitter dataset (Eisenstein et al., 2010).

Several researches have modeled the relationship between geographic location and language usage. Cheng, Caverlee and Lee (2010) suppose that tweets are noisy, and tweeters may have non-unique interest regions and positions. Thus, the "local words" associated with geographic centers and specific regions are useful for predicting tweeters' location (pp 759–768). Eisenstein, O'Connor, Smith, and Xing (2010) proposed a multi-level generative model to reason geographically-aligned lexical variation. And they use location and topic generate latent variables (pp 1277-1287). Pavalanathan, U. and Eisenstein, J. (2015) analyze the interaction between demographic variables, geolocation, and language usage. They also measured the usage rate bias associated with sampling techniques (pp 2138– 2148). Rahimi, Cohn, and Baldwin (2018) proposed a multi-view graphical neural network model which uses both text and graph information. And they found the neighborhood smoothing controlled by high-way network gates is essential to improving performance (pp 2009-2019).

3 Overview of word count dataset

The word count dataset mainly used for this study is uneven and sparse. As shown in figure 1, only one fourth of tweets in dataset were sent in west and mid-west region, while the two major classes 'south' and 'northeast' occupy the rest. In terms of users, they have similar region distribution as their tweets. Their tweet numbers are also similar regardless different regions, which is approximately long-tailed distribution with peak at around 30~50 tweets per user. Regarding the sparsity, an individual tweeter uses only around 6.7% of words in the vocabulary list on average, which is another challenge to models and feature engineering methods.

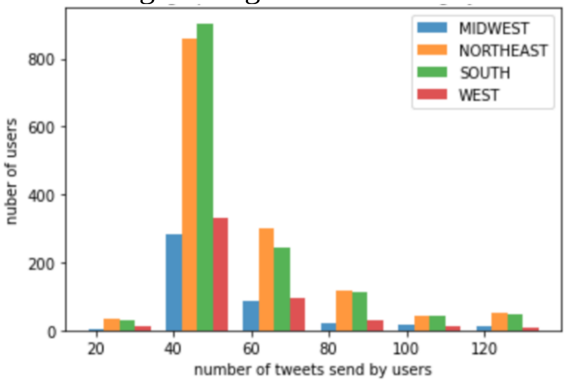

Figure 1- The distribution of the tweet num of user

4 Feature engineering

4.1 Feature selection

4.1.1 Method

This report examined two feature selection methods, then compared and analyzed their performance. One method is the blend feature selection method basing on the chi-square value and feature weight learned by xgboost. While another is manual feature selection and merging based on semantic.

The blend method selects two parts of features through the xgboost model and filter method, then uses their union as the final feature set. The first part was selected by xgboost's training results. The model was trained multiple times with different random seeds to choose features that have non-zero weight after training. The second part of the feature set was the top k features which have high correlation with regions measured by chi-square.

The manual feature selection is a somatic-driven attempt that tries to merge synonyms and discard meaningless words. To specific, it can be observed from the vocabulary list that multiple slangs have the same meaning, such as *haha* and *hahaha*. Besides, some words don't seem to have somatic meaning or relevant to the location, such as *16* and *â*.

4.1.2 Results and analysis

The blend feature selection improves the performance of model slightly, while the manual feature selection seems to have the reverse result.

model	Accuracy improves
Bayes	0%
LR	0.5%
Xgboost	3.7%
Random forest	5.2%
MLP	9%

Table 1- The accuracy improvement of different models after feature selecting.

As shown in table 1, this feature selection method improves most model's performance examined in this report by various degree.

To analysis the logic of feature selection, this report compared the words with high weight or measurement scores selected by mutual information, chi-square, and xgboost. It can be observed that the common-chosen ones are mainly region-related lexical variations and slangs. To be extract, the top 5 words selected by those methods are *#inhighschool, lmaoo, deadass,lls, and wassup,* which justifies the conclusion above.

Another phenomenon worth to notice is the performance of blend feature selection outperforms both xgboost and chi-square method, although it didn't always have the higher performance on individual models, as shown in table 2.

Feature selection method	Average improves on models
Blend	3.67%
Xgboost	-0.3%
Chi-square	3.23%

Table 2- The mean accuracy improvement of different feature selecting methods.

However, after removing seemingly meaningless words and merging synonyms to the same feature, the accuracy has decreased by 1.35% on average. That suggests the dataset was harder instead of easier to categorize after manual selecting. One possible reason is the usage of those slangs reflects the word usage preference of user from different regions.

4.2 Data preprocessing

4.2.1 Methods

The word count of different tweet send by the same user is averaged to create a steadier and condenser dataset. It can be assumed that all the tweets from the same user are generated by the same hidden variables related to that user, hence merging them into a single instance is reasonable.

4.2.2 Results and analysis

As shown in table 3, using the average word frequency of a user's tweets as instances not only reduces the size of the dataset but also improves the model performance significantly.

method	Accuracy	
	bayes	xgboost
Merge tweets	0.556	0.616
No process	0.458	0.455
Majority voting	0.336	0.341

Table 3- The accuracy of bayes and xgboost model in different pre-processed dataset

Compared with tweet-location prediction, the user-location prediction allows the model to learn the relation of tweet and user in a more macroscopic way, it may also reduce the noise by averaging. That may be the underlying reason of increase in perform.

The following check experiment justifies that the improvement of performance dues to the global view of users, rather than merely use the priory knowledge that the tweets from the same user must come from the same region. In this checking, the original counting data set is used for training and prediction, and the prediction results from the same user are unified by majority voting. As the table 3 shows, that method actually decreases the performance significantly instead of improving it to the same extent as merging users.

5 Model selection and evaluation

5.1 Baseline model

5.1.1 Method

Apart from the 0R model, the multinomial naive bayes is also chosen as the baseline model, because it is a simple linear model and often used for text classification.

5.1.2 Evaluation

As shown in figure 2, Naive Bayes has high deviation and low variance, and the model converges at a low number of instances.

Compared with the bayes model, the logistic regression outperforms the former by around 10% but often fails to converge. Probably because the assumption of independent features reduces the model to a determined set of underlying parameters. In contrast, the logistic regression tries to directly fit the data, which has difficulty in fitting the non-linear separable dataset.

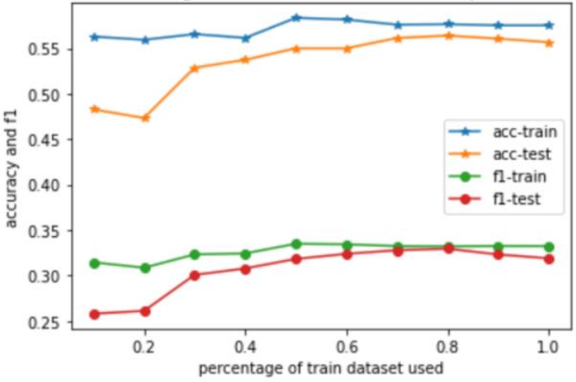

Figure 2- The learning curve of multinomial bayes model

5.2 Xgboost model

5.2.1 Method

The xgboost model was chosen as one machine learning model highlighted in this report for its high performance and interpretability. In addition, it can automatically fit non-liner data without specifying kernel function beforehand. Regarding the refinement, the greedy parameter turning strategy was applied due to the lack of computation resources.

5.2.2 Evaluation

In terms of performance, the xgboost model reached accuracy around 0.63 and f1 score around 0.5, higher than other model tested in this report.

During the parameter tuning process, the parameters seem to have a limited influence on generalization error. However, the greedy grid search does suggest the model tends to overfit and fewer trees with lower subsample proportion and less max depth may have a better performance. Those parameters suggest the need to balance variant with bias to decrease generalization error.

It can be concluded that the dataset is insufficient for the fit of xgboost. The figure 3 shows the xgboost with default model can perfectly fit all the data. And the trend of accuracy curve in figure 4 suggests the model didn’t converge with all the merged user as instance, and additional data may useful to improve the performance significantly.

Another phenomenon worth to mention is the bagging method with turned xgboost classfiers fails to decrease the variance, as shown in figure 5. That suggests the xgboost model learns the similar trees with different data samples, probably because the regulation parameters.

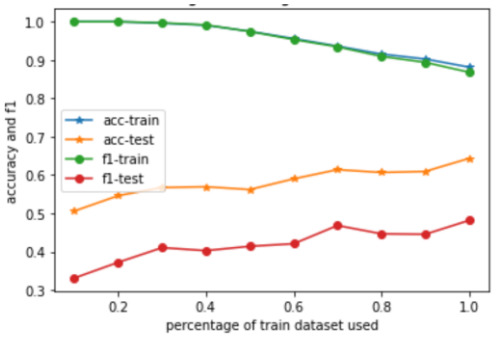

Figure 3- The learning curve of turned xgboost model

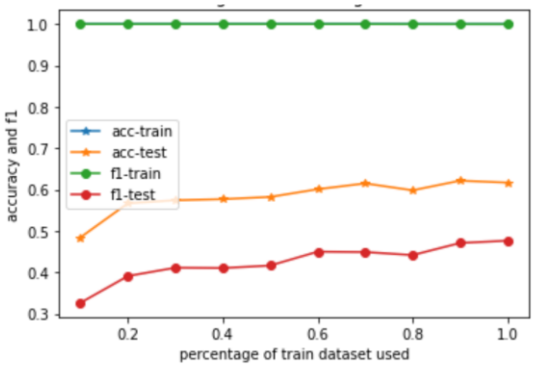

Figure 4- The learning curve of xgboost model

with default parameters

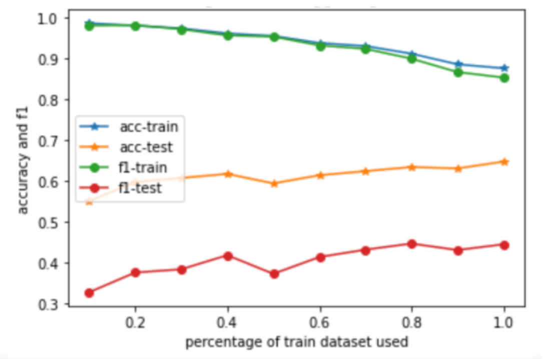

Figure 5- The learning curve of bagged xgboost model

5.3 Pre-trained bert classifier

5.3.1 Method

The pre-trained BertForSequenceClassification model was also examined on tweet text dataset in this report. That model consists of a pre-trained bert model and an output network which transform the output of bert model into classify label. The reason for choosing this model is its usual high performance on text classification tasks.

5.3.2 Evaluation

However, the performance of bert model is significantly lower than expected. To specific, its accuracy is 0.36, similar to 0R baseline.

One possible reason responsible for the bert model's poor performance may be the individual slangs and local words provide more information than the words sequence and semantic information in tweets. Those attention layers in the bert network flatten the region-related lexical variation, which is vital to classification. That results in the prediction of output network close to random guess between to majority labels. That can be justified by the training loss scatter. As figure 6 shows, the bert classification model has a high, unstable training loss which higher than 1 throughout the training process, which suggests the output network is unable to learn the information converted by the bert model.

6 Discussion

6.1 Class imbalance and moral issues

The accuracy and f1 score on regions with a larger population are significantly higher than the ones with a lower population. To specific, no user is classified as midwest by many models such as bert and naïve bayes. And most instances in minor regions like west and midwest are misclassified as south by nearly all the classifiers.

That bias shows the violation of the principle of fairness of models. Ethically speaking, every user should be treated equally by machine learning systems, including the geolocation classifier discussed in this report. However, fairness may conflict with performance in an uneven dataset. For example, the attempt to balance the dataset by subsampling discards around 50% of the dataset. Another instance of this conflict is the parameter turning process of SVC model, where assuming class weights results in a performance improvement of about 2%

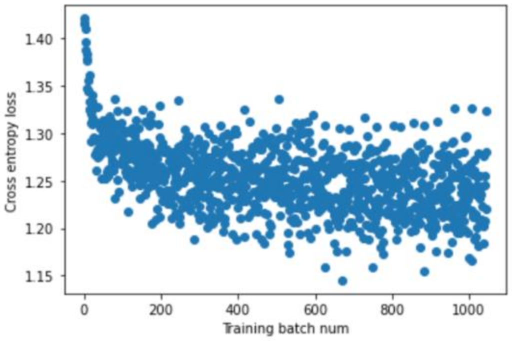

Figure 6- The loss of bert classifier during the training process

6.2 The major deficiency to improve

The main deficiency of this report is neglecting the relationship between users. As suggested by Cheng (2010) and Rahimi (2018), the neighborhood-based smoothing and social network may provide more information to improve the precision of classifying. However, the user information is only used as an identifier to distinguish tweets from the same user in this report. In addition, the logistic regression, which is liner model, have the similar low performance with more complex ones such as xgboost and multilayered perception on counts dataset. That may suggest the word frequency itself have limited ability for classify, and fully use the user information may provide additional support for classification.

In other words, it is worth to try the neighborhood-based smooth or social network related method in the future study.

7 Conclusions

In conclusion, this report shows the feature selecting, data preprocessing and model selecting methods and evaluated them afterward. The feature selection result and the performance of bert model suggest the top features useful to classifying are region-related lexical variations. Regarding the usage of user information, averaging the word frequency of the same user helps improve the performance.

References

Cheng, Z., Caverlee, J., and Lee, K. (2010). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 759–768.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. In Proceedings of the 2010 conference on empirical methods in natural language processing, pages 1277-1287.

Pavalanathan, U. and Eisenstein, J. (2015). Confounds and consequences in geotagged twitter data. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2138– 2148.

Rahimi, A., Cohn, T., and Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2009–2019.